Journal de la Société Française de Statistique
Vol. 160 No. 1 (2019)

Bootstrap Validation of the Estimated Parameters in Mixture Models Used for Clustering

Titre: Validation par bootstrap de l'estimation des paramètres d'un modèle de mélange utilisé en classification

Zhivko Taushanov[1] and André Berchtold[2]

Abstract: When a mixture model is used to perform clustering, the uncertainty is related both to the choice of an optimal model (including the number of clusters) and to the estimation of the parameters. We discuss here the computation of confidence intervals using different bootstrap approaches, which either mix or separate the two kinds of uncertainty. In particular, we suggest two new approaches that rely to some degree on the model specification considered as optimal by the researcher, and that address specifically the uncertainty related to parameter estimation. These methods are especially useful for poorly separated data or complex models, where the selected solution is difficult to recreate in each bootstrap sample, and they present the advantage of reducing the well-known label-switching issue. Two simulation experiments based on the Hidden Mixture Transition Distribution model for the clustering of longitudinal data illustrate our proposed bootstrap approaches.

Résumé : Lorsqu'un modèle de mélange est utilisé en classification, l'incertitude est liée au choix du modèle optimal (y compris le nombre de groupes) et à l'estimation de ses paramètres. Nous discutons ici du calcul d'intervalles de confiance en utilisant différentes approches bootstrap qui mélangent ou au contraire séparent ces deux types d'incertitude. En particulier, nous suggérons deux nouvelles approches qui dépendent en partie de la spécification du modèle considéré comme optimal par le chercheur, et qui répondent spécifiquement à l'incertitude liée à l'estimation des paramètres. Ces méthodes sont spécialement utiles lorsque les données sont mal séparées ou lorsque le modèle à estimer est complexe et que la solution choisie se révèle difficile à reproduire dans chaque échantillon bootstrap. De plus, elles présentent l'avantage de réduire le problème du label-switching. Deux simulations basées sur le modèle *Hidden Mixture Transition Distribution* adapté à la classification de données longitudinales illustrent nos propositions.

Keywords: clustering, mixture model, bootstrap, uncertainty, label-switching, confidence interval, frequentist estimation, HMTD model
Mots-clés : classification, modèle de mélange, bootstrap, incertitude, label-switching, intervalle de confiance, estimation fréquentiste, modèle HMTD
AMS 2000 subject classifications: 62F40, 62H30

1. Introduction

This study discusses different bootstrap approaches to investigate the parameter uncertainty in mixture models used for clustering, that is for grouping similar data points together. In the frequentist estimation of mixture models, each bootstrap sample has its own permutation of the mixture components. This causes the so-called label-switching problem, which is very well-known in the Bayesian context, but rather neglected in the frequentist case, because it does not occur

[1] University of Lausanne, Institute of Social Sciences & NCCR LIVES, Switzerland
E-mail: jivko.taushanov@gmail.com
[2] University of Lausanne, Institute of Social Sciences & NCCR LIVES, Switzerland
E-mail: andre.berchtold@unil.ch

 ISSN: 2102-6238

when only a point estimate of the parameters is required. Here, we propose and compare different strategies to overcome this issue and to evaluate the disadvantages involved.

In all clustering applications, a major concern is the uncertainty of the obtained solution; that is, how sure one can be that the obtained clustering reflects the true structure of the population. We also need to evaluate whether the estimated parameters are significant, and hence useful, or not, and how they would vary if another sample was drawn from the same population. Finally, we also need to identify the specific characteristics of each cluster to distinguish the clusters from each other. In this study, we assume that one uses a representative sample of the population under study, and we separate the remaining uncertainty in two parts: one related to the choice of the model and of its specification, including the number of components, and one related to the variability of the estimated parameters once the model has been specified and approved.

In the frequentist framework for mixture models, parameter inference may be performed using bootstrap. The standard way is to resample from the entire sample before fitting the clustering model (*non-parametric bootstrap*). By doing this, we mix the parameter and model uncertainties and this can complicate the task much more due to issues such as the label-switching problem. Therefore, in this article, we introduce two alternative bootstrap procedures: The first possibility is to isolate each type of uncertainty by separating the sample according to an already accepted clustering partition (*separate bootstrap*). The second, hybrid possibility is to assume that the clustering obtained from the first optimization of the model is correct and to draw stratified samples with respect to the accepted solution (*stratified bootstrap*), but again with the risk of mixing both types of uncertainty.

In the remaining of the paper, we begin by describing the use of mixture models for clustering and present the well-known label-switching problem as well as other related issues. Then, we discuss parameter inference and describe in detail the two new types of bootstrap (*separate* and *stratified*). Two simulation experiments are provided to demonstrate the usefulness of the new types of bootstrap and the paper ends with a discussion.

2. Mixture models for clustering

2.1. Validation of a clustering

A mixture model is a statistical model combining several different possible generating mechanisms called *components* for the data under study. When used for clustering, each data point is then assumed to have been generated by one of these components, and the goal is then to assign each data point to the component that generated it, hence creating a clustering of the dataset in k mutually excluding groups. Once a solution has been found, it is then necessary to validate it. This includes two dimensions: finding an optimal clustering in terms of model choice and number of clusters on one hand, and inference on the model parameters on the other hand. In this study, we separate these two dimensions and concentrate exclusively on the latter.

Mixture models entail specific problems, the best-known being the *label-switching problem* described in the next section, but more generally, they differ from other statistical models for the following reasons: They belong to the so-called *ill-posed problems* family because they do not satisfy all three properties of a well-posed mathematical problem (i.e., a solution exists, the solution is unique, and the behavior of the solution changes continuously with the initial conditions). More

 ISSN: 2102-6238

precisely, mixture models can have multiple solutions and, most importantly, small changes in the data can have a large impact on the results. Moreover, two different parameterizations can sometimes lead to similar joint distributions, meaning that a unique solution does not necessarily exist. This is also an *inversed problem*, because the data provide information on the parameters indirectly: we extract information on the data-generating process from the data itself. Thus, there actually exists a non-null probability for a component of the model to be empty, and, in such a case, the sample cannot provide any information about its parameters. This can explain why the likelihood function can become unbounded (Marin, Mengersen & Robert, 2005).

2.2. The label-switching problem

An important issue in parameter inference arises from the fact that a likelihood function is invariant to a permutation of its components. This implies that each solution of a model can have the same components, but in a different order. This is the so-called *label-switching problem*, which is common in mixture models and especially important when assessing parameter uncertainty. The optimal solution in mixture models is typically searched for by maximizing the likelihood but, depending on the model complexity, it can be difficult to find an optimal solution, especially because the likelihood function can have multiple local optima. Although many mixture models are theoretically identifiable, in practice, obtaining an optimal solution can prove much more difficult, especially when working with different samples obtained from the same population, bootstrap samples for instance. Therefore, two supposedly similar joint density distributions may be estimated as two different mixtures and the problem is even more pronounced when the dataset is small. This situation aggravates the label-switching problem, and we cannot always find a correspondence between the labelling of the components in each solution.

Label-switching is a major issue in the Bayesian estimation of mixture models, and much research has been done in this field (Celeux, Hurn & Robert, 2000; Stephens, 2000; Jasra, Holmes & Stephens, 2005; Sperrin, Jaki & Wit, 2010). On the other hand, much less attention has been paid to this problem in the frequentist context, because if one is interested only in the density of the mixture (i.e., the point estimation of parameters), label switching does not occur. However, it also becomes an important issue in *frequentist* estimation when one needs to use a resampling procedure, such as bootstrap, to evaluate the variability of the parameters. The reason is that, in this type of mixture model estimation, each bootstrap sample has its own permutation of the mixture components, and the components have to be matched together, hence creating the need for methods that can identify the components.

An issue related to label-switching is the multimodality problem, that is, the presence of multiple modes in the distribution. If we have k components in a mixture model, the number of modes is of order $O(k!)$, as we must consider all possible permutations of the indices. Different authors (Celeux, Hurn & Robert, 2000; Rodriguez & Walker, 2014) stress that label-switching is necessary for the convergence of MCMC, since, without it, the sampler is not exploring all the mixture model's $k!$ possible posterior distribution modes (leading to poor sampler mixing). However, as will be seen later, multimodality can also constitute a problem when using the bootstrap.

Several strategies to solve the label-switching issue have been proposed. The first one is to introduce *identifiability constraints* to remove the symmetry of the likelihood, but this strategy

 ISSN: 2102-6238

has been shown to fail in some cases (Stephens, 2000). Various *relabeling approaches* (Stephens, 2000; Celeux, Hurn & Robert, 2000) have also been presented. In frequentist estimation, random initialization is sometimes sacrificed to avoid label-switching (O'Hagan, Murphy, Scrucca & Gormley, 2018). This approach appears sufficient for simple specifications of mixture models, but it does not guarantee effective results for more complex models or for less separated data, and much research is required to validate this approach in the general case. In general, none of the proposed methods can be considered as always performing well, especially for complex frequentist mixture models with several parameters per component or with a high number of components. In such cases, avoiding to deal with the label-switching issue could be a better solution, and this is the aim of the alternative bootstrap approaches presented in the next section.

3. Parameter inference in mixture models

3.1. Existing methods

In real-world clustering problems, one often has no a priori knowledge of the underlying groups. However, through inference on mixture models, one may try to recover the latent cluster membership of the observed data, to provide an estimation of the parameters describing the characteristics of the different groups, and to find the optimal number of groups. Therefore, one needs to obtain an optimal clustering solution (in terms of model-selection criteria), but also to explore the variability of the estimated parameters and to assess the stability of the underlying clusters. The different objectives mentioned above reflect different sources of uncertainty, but we will now focus on the assessment of the significance and of the dispersion of the estimated parameters through the computation of confidence intervals (CIs).

Several approaches to compute CIs in the case of mixture models have been proposed, but, depending on the type and the complexity of the models, not all of the corresponding methods are feasible. We summarize below some well-known methods:

— **Finite-difference approximation of the Hessian matrix** is the most logical choice for estimating parameters and their standard errors since the maximum likelihood estimator (MLE) is asymptotically normal. As the sample size n increases, we have $\hat{\theta}_{ML} \xrightarrow[n\to\infty]{} \mathcal{N}(\theta, \frac{\mathscr{I}^{-1}}{n})$, where $\mathscr{I}$ is the Fisher information matrix evaluated at the true value of parameter θ. $\mathscr{I}$ can be estimated from the data as the Hessian of the log-likelihood evaluated at $\hat{\theta}_{ML}$. Then, one can easily estimate a CI for any parameter as follows:

$$\hat{\theta}_{i,j}^{MLE} \pm \frac{z_{1-\frac{\alpha}{2}} \left(\hat{\mathscr{I}}^{-1}\right)_{ii}}{\sqrt{n}}$$

As stated by Rydén (2008), using the covariances and the fact that the sum of squared standard Gaussian distributions results in a $\chi^2(q)$ distribution, one may obtain "ellipsoid" CIs for the parameters, which are given by the region defined by $(\theta - \theta_{MLE})^T \hat{\mathscr{I}} (\theta - \theta_{MLE}) \leq \chi^2_{1-\alpha}(q)$. Notice also that the asymptotic normality assumption is very important for such CI computations, and that they can become especially problematic when the parameters are clearly non-Gaussian, such as in the case of variances, for instance.
Moreover, one may encounter problems when estimating the Hessian (Visser, Raijmakers & Molenaar (2000)), which is especially true in the case of the Hidden Mixture Transition

 ISSN: 2102-6238

Distribution (HMTD) model used in the simulation section of this article. Indeed, derivatives are very difficult to obtain for this model, because of the complexity of the log-likelihood equation when the variances are not constant.

— **Likelihood profiles** consist of an expansion of the likelihood function around the maximum likelihood (ML) estimate of each parameter separately (Visser, Raijmakers & Molenaar, 2000). Assume that we have a two-parameter model with parameters ϕ and θ. If we are interested in ϕ, then θ is treated as a nuisance parameter to obtain the likelihood profile

$$L_p(\phi) = \max_{\theta} L(\phi, \theta).$$

The value of ϕ that maximizes this function is denoted by ϕ_M. Then, ϕ is moved away from its maximum, and using the new value of ϕ_0, the following ratio is re-estimated:

$$R_p = -2(log(L_p(\phi_M)) - log(L_p(\phi_0))) = -2\ln\frac{L_p(\phi_M)}{L_p(\phi_0)}.$$

The procedure is repeated until the value of ϕ_0 corresponding to $R_p = 3.841$ is found. The latter value is the threshold of the $\chi^2(1)$ distribution corresponding to $\alpha = 5\%$. This represents finding the limits ϕ_0 of the region beyond which the likelihood ratio test becomes significant (i.e., the null model becomes significantly different from the optimal one). The CIs are obtained by repeating the same procedure on both sides of the ML estimate.

These first two methods assume asymptotic normality, but the three following alternatives based on resampling do not have this requirement. One of the main motivations for using resampling approaches is that the complexity of the likelihood function in some mixture models makes it very difficult to obtain the Fisher information matrix.

— The **non-parametric bootstrap** (Efron, 1979) is a versatile tool of modern statistics. One problem in clustering is to eliminate, as much as possible, the sampling error stemming from working with a sample instead of with the entire population. If the sample size is very large, one may divide it and apply clustering independently on the different subsamples, but since the sample size if often limited in practice, the only way to approximate the true underlying distribution, F, of the data is through the empirical distribution $\hat{F}_n$ of the sample. The bootstrap consists of building samples by drawing observations at random and with replacement from the original sample. The statistical model of interest is then fitted to each of the resulting samples to obtain a corresponding number of parameter estimates. The CIs are then directly obtained from the quantiles of the distribution of estimates. In frequentist mixture models, the bootstrap is the main tool used for parameter inference.

— The **jackknife** has also been used to estimate parameter variability for hidden Markov and mixture models (O'Hagan, Murphy, Scrucca & Gormley, 2018). As an old predecessor of the popular non-parametric bootstrap, this method is similar to the latter, except that the subsamples are obtained by removing one observation at a time from the original sample. Regarding Markov models, the jackknife has not shown better results than the bootstrap, as demonstrated by Rosychuk, Sheng & Stuber (2006). That explains why this method has often been replaced by the non-parametric bootstrap since its introduction by Efron.

— The **parametric bootstrap** uses the original sample to obtain ML estimates for a given model. The estimated model is then used to generate a simulated sample of the same size as

 ISSN: 2102-6238

the original sample. The model is fitted to the simulated sample and the model parameters are estimated again. The procedure is repeated n times to generate an empirical distribution around the ML estimates of the parameters. The most obvious problem with parametric bootstrap is the assumption that the model can perfectly explain all the features of the original sample.

Non-parametric bootstrap has been used previously to compute CIs in mixture models. In combination with label-switching solutions, this method has shown positive results for simple mixtures. Grün & Leisch (2004) showed that the bootstrap is useful for evaluating the stability of parameters when estimating finite mixture models and they recommend it in addition to multiple initialization of the EM algorithm. Using simulated data, the authors apply both a parametric and non-parametric bootstrap, to find that the parametric bootstrap is also a useful tool for analyzing the stability of parameter estimates. As a gold standard, we will focus on the non-parametric bootstrap to compute CIs. However, this technique does not guarantee that the underlying clusters will be represented equally in each bootstrap iteration. Combined with the multimodality issue, this may sometimes lead to incompatible clustering solutions, necessarily introducing a bias in the CIs and, therefore, in the conclusions regarding parameter significance as well. The problem is even more pronounced when small samples are used, or when one of the clusters is small. Finally, the label-switching issue can become difficult to solve when working with rather complex mixture models with multiple components or multiple parameters for each component. Therefore, the best solution is perhaps to avoid these issues rather than to have to deal with them. This is our motivation for looking at different bootstrap methods for parameter validation. We propose, hereafter, two possible alternatives to the non-parametric bootstrap and discuss their advantages and disadvantages.

3.2. Separate bootstrap

Examples of bootstrap procedures for clustering models, such as the ones proposed by Grün & Leisch (2004), are based on very simple two-component mixtures, for which the application of simple identifiability constraints is sufficient. The use of the bootstrap in more complex models appears much more difficult. Our proposal for such problems is to apply bootstrap estimation only after a reliable valid solution (cluster partition) is found, what we will call the *chosen solution* from now. This is equivalent to saying that once an acceptable solution has been found, each cluster can be considered as a mutually independent population. Then, the bootstrap procedure is performed separately on each cluster of the chosen solution, with the CIs computed using a single-component model for each generated sample.

Indeed, this implies that we consider the chosen solution as the best that can be found, and neglect the fact that another may exist, possibly yielding a better fit to the data. However, a better fit does not necessarily imply a better and more useful clustering. Therefore, one needs to find the solution most suitable to the data based on all available information, including that from statistical criteria and from knowledge of the data. In this case, by applying the bootstrap to the chosen clustering partition, we do not measure the stability of the solution, but rather isolate the variability of the parameters for this particular clustering partition. This makes sense, especially when two bootstrap iterations of a given complex model may be completely incompatible (due to cluster instability, for instance), resulting in the relabeling strategy being proven wrong.

 ISSN: 2102-6238

Another advantage of this approach is that we avoid the typical singularity problem in likelihood maximization, because we use a single-component model for constructing the intervals.

This procedure may also be effective in "soft" clustering. For instance, in time series, we are often working with a single, very long data sequence. By bootstrapping the part of the time series associated with each hidden state separately, we could be able to interpret the different components of the model more accurately and with more certainty.

As a limitation, note that the parameters for the components' weights are, indeed, not inferable through separate bootstraps, but must be assessed beforehand, while choosing the optimal model solution. On the other hand, the separate bootstrap has the advantage of completely eliminating the multimodality issue. Single-component models imply that we can avoid having another mixture solution that gives similar joint density results. Furthermore, the label-switching and singularity problems are alleviated.

3.3. Stratified bootstrap

One reason for the difficulty in correctly identifying the clusters obtained from a bootstrap sample is the difference between the proportion of data in each cluster in the original sample and in the bootstrap samples. This is especially so when the proportions of data in each cluster are very different from one another. One possible solution is to include the defined cluster proportions in the bootstrap procedure and ensure a proportional presence of data from all the presumed clusters in each sample. Now, assume that our original sample is subjected to a clustering procedure with c classes. The obtained solution assigns on of the class labels $\omega_j \; j \in [1, ..., c]$ to each observation. These class labels are then used for a bootstrap on the original sample, but by using a sampling procedure that respects the chosen clusters' proportions at each iteration. In other words, we create the bootstrap samples by selecting from each cluster a portion of data proportional to the cluster's size in the chosen clustering partition. The full clustering model is then applied on the bootstrap sample. This approach may be considered as a kind of "stratified" bootstrap, where the "strata" are defined by a model solution (partition), obtained from the original sample before performing the bootstrap.

The advantage of this procedure is that it maintains the proportions of the already validated clustering obtained from the original sample. This may result in more stable bootstrap estimates for the parameters compared to those by the ordinary non-parametric bootstrap sampling of the original data. The effect is particularly important in small samples (or in large samples with small but clearly distinct clusters) because, in the basic non-parametric bootstrap on a small sample, the representativeness of each possible underlying class in the data is not respected, probably leading to the clustering of the bootstrap sample finding a completely different solution compared to that of the original sample. In other words, very small but distinct classes may "vanish" in resampling (for example, voters of small political parties in a survey, or low-probability components in a mixture). In this case, all the relabeling methods will prove useless, because the components of each solution would simply be incompatible with each other. Of course, our approach will not eliminate the label-switching issue, but it can reduce it by increasing the probability of finding bootstrap solutions close to the original solution.

Compared to the separate bootstrap presented above, the chosen solution in a stratified bootstrap has less influence on the final results because the elements of each class are still represented and

 ISSN: 2102-6238

can still be drawn during the resampling procedure.

In a recent study, O'Hagan, Murphy, Scrucca & Gormley (2018) tested different bootstrap procedures in mixture models and proposed an interesting alternative to the stratified bootstrap, namely, the weighted likelihood bootstrap (WLBS), using the R package mclust (Scrucca, Fop, Murphy & Raftery , 2016). At each bootstrap iteration, WLBS uses the original sample, but with different weights, simulated from a Dirichlet distribution. Compared to the stratified bootstrap, this approach is less influenced by the chosen solution, but it could be less effective when the data are not clearly separated in well-defined groups and we want to repeatedly find the same partition more easily (see Berchtold, Suris, Meyer & Taushanov (2018) for a practical example).

4. Simulation study

In this section, we present two simulation experiments designed to provide evidence of the respective behaviors of the non-parametric, separate, and stratified bootstraps. These simulations use longitudinal data and a mixture model adapted to this case, the Hidden Mixture Transition Distribution (HMTD) model.

4.1. The HMTD model for longitudinal data

The HMTD model is a model able to describe and cluster sequences of continuous longitudinal data (Bolano & Berchtold, 2016). It combines a visible level and a latent level. The visible level is a Mixture Transition Distribution model first introduced by Raftery as an approximation of high-order Markov chains (Raftery, 1985) and then developed by Berchtold & Raftery (2002) and Berchtold (2003). Here, we use a Gaussian version of the model, where the mean of the Gaussian distribution is a function of past observations:

$$\mu_{g,t} = \varphi_{g,0} + \sum_{i=1}^{p_g} \varphi_{g,i} x_{t-i}$$

where $\varphi_{g,0}$ is the constant for the mean of component g, $\varphi_{g,i}$ is the autoregressive parameter indicating the link between x_{t-i} and x_t, and p_g is the number of lags. In our simulations, the variance of each component is constant, but it could also be written as a function of past observations.

The latent level of the HMTD model is a homogeneous Markov chain. Each state of the chain is associated with a different Gaussian component at the visible level, with the transition matrix of the Markov chain being used to determine which component best represents the current observation. To use the HMTD model as a clustering tool, we assume the hidden transition matrix to be the identity matrix. Consequently, each sequence of observations is associated with only one component of the model, thus generating a clustering of sequences into mutually exclusive groups. The HMTD model is estimated by maximizing its log-likelihood through a generalized expectation maximization algorithm (Taushanov and Berchtold, 2017a)

4.2. Experiment 1

In this experiment, we compare the behavior of the three bootstrap procedures (non-parametric, separate, and stratified) for the computation of CIs for the parameters of an HMTD model, applied

 ISSN: 2102-6238

to a dataset consisting of 150 sequences, each of length 25, generated (after a burn-in period of 25 data points) by one of the following two AR processes: One hundred sequences were generated from the AR(2) process

$$x_t = 2.5 + 0.4 \times x_{t-1} + 0.3 \times x_{t-2} + \varepsilon_1; \quad \varepsilon_1 \sim \mathcal{N}(0, 2^2)$$

and 50 sequences were generated from the AR(1) process

$$x_t = 0 + 0.9 \times x_{t-1} + 0 \times x_{t-2} + \varepsilon_2; \qquad \varepsilon_2 \sim \mathcal{N}(0, 2^2)$$

Thus, the true values of the parameters that we attempt to estimate are $\varphi_{01} = 2.5, \varphi_{02} = 0, \varphi_{11} = 0.4, \varphi_{12} = 0.9, \varphi_{21} = 0.3, \varphi_{22} = 0, \theta_1 = 2$, and $\theta_2 = 2$. Notice that the coefficient φ_{22} is not required, since its value is zero, but we chose to keep it, to check that it is correctly estimated to zero by the optimization algorithm and by the bootstrap procedure.

The chosen model is the first two-component solution that we obtained. In this chosen clustering solution, a small number of the sequences were misidentified: nine sequences were wrongly assigned to the second group, whereas three were misclassified in the first group. This small misclassification in the chosen separation into clusters could introduce some bias in the CIs from the methods that use the chosen solution ("separate" and "stratified"). For example, the nine misclassified sequences originally generated from component 1 have a non-zero probability of being included in the bootstrap samples used to estimate the parameters of component 2; these estimates will therefore be biased towards the value of the parameters of component 1. However, this is similar to what can be expected in real situations. Moreover, since the proportion of misidentifications is small, the CIs should still be able to recover the true parameter values. Table 1 summarizes the 95% CIs obtained from the quantiles of the bootstrap distribution after 300 bootstrap replications.

TABLE 1. *Experiment 1: 95% CIs obtained for the HMTD model parameters using the three types of bootstrap.*

First component	σ_1^2	φ_{01}	φ_{11}	φ_{21}
True values	**4**	**2.5**	**0.4**	**0.3**
Non-parametric bootstrap	(3.633; 4.215)	(2.076; 2.915)	(0.339; 0.422)	(0.276; 0.362)
Separate bootstrap	(3.645; 4.100)	(2.261; 2.830)	(0.341; 0.422)	(0.275; 0.357)
Stratified bootstrap	(3.640; 4.206)	(2.084; 2.870)	(0.331; 0.422)	(0.276; 0.375)
Second component	σ_2^2	φ_{02}	φ_{12}	φ_{22}
True values	**4**	**0**	**0.9**	**0**
Non-parametric bootstrap	(3.746; 4.406)	(-0.022; 0.272)	(0.839; 0.983)	(-0.060; 0.123)
Separate bootstrap	(3.808; 4.380)	(-0.023; 0.178)	(0.895; 0.997)	(-0.062; 0.053)
Stratified bootstrap	(3.721; 4.461)	(-0.011; 0.232)	(0.841; 0.989)	(-0.060; 0.113)

As expected, all the CIs effectively recover the true values of the parameters. The separate bootstrap provided a systematically narrower CI compared to the ones from the other two approaches, but this was as expected, because the uncertainty is lower when we only have to estimate the parameters of a one-component model. On the other hand, the non-parametric and stratified bootstraps yielded similar results, certainly because the model was simple enough to be easily estimated with the ordinary procedure. In more complicated real-world situations, for instance, with more unequal cluster sizes, or with poorly separated clusters, a difference should appear, with a wider CI from the non-parametric bootstrap compared to that from the stratified procedure.

 ISSN: 2102-6238

However, in such situations, it would become more difficult to compare the three approaches, because of the increasing complexity of the label-switching issue and the difficulty to detect irregular solutions from the stratified and non-parametric bootstraps.

The label-switching issue arose when we attempted to relabel the ordinary and stratified bootstrap solutions. However, since our example has only two components and, although the sequences from the two generating processes were overlapping, we easily solved the problem by comparing the group memberships of the chosen model solution and the bootstrap solutions, and by selecting the label with the best match. However, the complexity of the likelihood function caused some bootstrap solutions to become stuck in local optima and exhibit irregularities, the most common one being convergence to a one-cluster solution. We need to identify such degenerated solutions and remove them from the CI calculation to avoid additional bias. One obvious way to discover these solutions is to check for the presence of all the components. Irregularities are, however, not limited to the absence of one component. A local optimum can be a solution incompatible with the chosen solution even when all components are used. For instance, highly influential or extreme sequences may be drawn several times in the same subsample (especially when the sample size is small), creating their own component. A numerical likelihood optimization problem can also lead to aberrant solutions. Therefore, after the relabeling procedure, we need to check for the presence of extreme values in the parameters of each component. In our experiment, 19 out of 300 solutions from the stratified bootstrap were found to be irregular, and so were 20 solutions from the ordinary non-parametric bootstrap. The separate bootstrap was not influenced because all calculations were made separately for both of the components in the chosen solution.

4.3. Experiment 2

The second simulation experiment extends the previous results by comparing the behavior of the three bootstrap procedures as a function of the length of each data sequence, and of the number of sequences. It also uses a more complex three-component model with fixed variances and one or two lags per component (refer to Table 2 for the true value of each parameter). Notice that the comparison of the results from datasets with a different number of sequences and sequence lengths comes at a cost: A different initial dataset has to be used for each situation, and, hence, the results are not fully comparable between situations. Therefore, this second experiment used a setting slightly different from the one of the pure bootstrap in Experiment 1. Three situations were considered: Situation 1, with 200 sequences of length 25; situation 2, with 200 sequences of length 100; and, situation 3, with 800 sequences of length 25. In each case, half of the sequences were generated from component 1, 25% of them from component 2, and 25% of them from component 3. In contrast to Experiment 1, here we did not rely on a unique dataset from which bootstrap samples were created, but we generated a different dataset for each replication and each situation. The advantage is that the models estimated in each replication are less influenced by the peculiarities of a unique dataset as it could have occurred in Experiment 1. By using a different dataset for each replication, we suppress as far as possible the influence of a specific dataset upon the final results of the experiment, the observed differences being then truly attributable to the different numbers and lengths of the sequences. However, we acknowledge that this special setting makes Experiment 2 farther from a real case than Experiment 1. Table 2 summarizes the results

 ISSN: 2102-6238

of the second experiment using 800 replications in each of the three situations.

Since a different dataset was used in each replication, and since the model was more complicated than the one in Experiment 1, the CIs are also wider than those in Experiment 1, especially regarding the variances. However, the goal here is not to evaluate the performance of the HMTD model estimation procedure, but to evaluate the impact of changing the number or the length of sequences on the CIs. Compared to the basic situation (situation 1: 200 sequences of length 25), situations 2 and 3, which have more data points, usually provide narrower CIs, no matter the bootstrap procedure, but this is more pronounced in situation 2 with the increase in the length of sequences rather than in their number. This improvement was expected since, in general, more information leads to more accurate results, but this had to be verified in the case of the separate and stratified bootstraps. When comparing the three bootstrap procedures, the separate bootstrap, and to a smaller degree the non-parametric bootstrap, usually takes better advantage of the increase in the number of data points than the stratified bootstrap. This is logical, since the latter bootstrap procedure relies in part on the chosen "best" model, but is still influenced by the uncertainty related to the structure of the clustering model. On the other hand, the separate bootstrap simplifies the optimization of the model for each component, and the non-parametric bootstrap is less influenced by an incorrect proportion of the different components as the number of data points increases.

4.4. Main lessons

Experiment 1 was designed to be as close as possible to a real setting, with one model chosen as the "correct" solution by the researcher. On the other hand, Experiment 2 used a slightly different dataset, and hence estimated model, for each replication of the bootstrap procedures, with the goal of understanding the role of sample size in the computation of CIs by the three different bootstrap procedures. Overall, the two alternative bootstrap procedures introduced in this article provide results consistent with those of the traditional non-parametric procedure. The resulting CIs are centered around the same value and exhibit the same general shape, either being symmetrical, or skewed to the left or to the right. However, the fact that the separate and, to a lesser extent, the stratified bootstrap, is easier to compute, notably by limiting (in the case of the stratified bootstrap) or suppressing (in the case of the separate bootstrap) the label-switching problem, implies that these two procedures lead to CIs narrower than those obtained from the non-parametric bootstrap. We can hypothesize that, since the CIs obtained from the non-parametric bootstrap could be wider than required due to the errors implied by the label-switching problem, the CIs obtained from the alternative bootstrap procedures could, on the other hand, be too narrow. This calls for additional simulation experiments designed to determine the coverage level of the CIs obtained through each bootstrap approach.

5. Conclusion and discussion

Putting the sample representativeness aside, the actual cluster membership of each observation is unknown, and so is the number of clusters; in some cases, even any "real" separation between the data is absent. Uncertainty in clustering may be divided into several components, such as the

 ISSN: 2102-6238

presence and type of heterogeneity in the data, the number of clusters, the stability of the partition, and the significance of the parameters (which are characteristics of the clusters).

The presence of several levels of uncertainty makes it difficult to independently assess the variability of the estimates, because these estimates depend on the structure of the clustering model, such as the number of components and the variance-covariance matrix. For instance, if the number of clusters is incorrect, the cluster membership of a data point will also be incorrect, in turn rendering the parameter estimates incorrect as well. On the other hand, incorrect parameter estimates may lead to a different optimal number of clusters and erroneous cluster memberships, since clusters near each other may merge together and erratic parameter estimates often result in unlikely empty clusters. Therefore, model-based clustering would generally make it very difficult to cope with all the above problems simultaneously. This leads to the following question: When we are dealing with real-world poorly separated data, is it accurate to estimate the uncertainty of the cluster weights and to attempt to find the same model within every bootstrap sample, or is it better to find and approve one interpretable separation of the data, on which we can base our uncertainty estimation?

Various internal and external cluster validation methods and information criteria have been proposed to validate a clustering solution (Meila, 2016), but this is not the purpose of our study. Here, we explored the particularities and issues that arise when assessing the uncertainty of the parameter estimates of mixture models in the presence of label-switching and other problems. Two new types of bootstrap were presented, which reduce the impact of label-switching. They rely (to a different extend) on the best clustering solution approved by the researcher. Both methods, but especially the *separate* bootstrap, implement the idea of separating the model uncertainty from the parameter uncertainty. When using the ordinary non-parametric bootstrap, the tasks of discovering the optimal clustering model, validating it, and calculating the parameters' confidence intervals are performed simultaneously. On the other hand, the *stratified* bootstrap is a hybrid approach, which uses the chosen solution but allows for the possibility of a model choice error and minimizes its impact on the final results. This ensures that the chosen clusters are well-represented throughout the bootstrap sampling. In the *separate* bootstrap, the model obtained from the original sample is assumed to be the correct model (at least with this number of clusters and these parameters), so we only need to validate its parameters. Separate bootstrap procedures are then applied independently on each cluster defined by the approved solution.

When considering real-world problems, many datasets do not perfectly satisfy the underlying assumptions of parametric mixture models. Nevertheless, these models are still considered very useful in practice. From this point of view, the objective could not necessarily be to identify the "best" clustering model from a statistical point of view, but the most appropriate one, given the dataset and the research question. Choosing such a model is often the only way to explore how significant the different parameters are, given our speculations regarding their distribution. In contrast, mixing the model and parameter uncertainties often leads to inconclusive results in frequentist estimation, especially when the clusters are difficult to identify.

If we assume that a "best" clustering partition always exists, either from a statistical or from an interpretability point of view, the chosen solution may contain misclassified observations, which could introduce bias in the CI calculations. This is especially important for the separate bootstrap method, and this highlights the importance of the chosen solution. However, when the sequences are well-identified from the beginning and the chosen number of components is appropriate,

 ISSN: 2102-6238

this method can prove advantageous over its alternatives. Furthermore, model uncertainty is not always a primary objective in practice, because the researcher often attempts to find not only the stability of a clustering solution compared to all the other possible alternatives, but also an adequate clustering based on the knowledge of the data behavior and of the aim of the study. In this case, one may ask either "By how much could the characteristics of the solution vary?" or "Is this clustering the only one appropriate for those data?". Although a separate bootstrap is clearly more appropriate for answering the first question, the non-parametric and stratified bootstraps are better suited for answering the second one. Therefore, each procedure has its utility, depending on the problem to be solved. However, a major advantage of the separate bootstrap compared to the other procedures is the suppression of the label-switching, multimodality, and singularity issues.

In addition to the above considerations, and as shown by the second simulation experiment, the separate bootstrap may also be useful for small samples, or where at least one of the groups appears obvious but its representation in non-parametric bootstrap samples is uncertain. On the other hand, the separate bootstrap also presents some disadvantages, such as the impossibility of inference on the mixing weights and of imposing particular covariance structures on the mixture models (Celeux & Govaert, 1995). Nevertheless, the separate bootstrap may be the only possibility available for overcoming the label-switching issue in the frequentist estimation of mixture models, especially when the likelihood equation is very complicated due to a large number of components, non-Gaussian distributions, or distributions depending on multiple parameters.

To summarize, the most important decision faced by the researcher is to select the best clustering model based on the observed data (type of model, and model parameters, including the number of components). Then, using a suitable estimation procedure, one possible solution will be found and evaluated using statistical criteria and possibly other information. Then, three situations are possible:

1. If a stable, interpretable solution is validated by the researcher, then the separate bootstrap can be used to compute confidence intervals. The advantage is that the label-switching problem is eliminated, as well as the multimodality and singularity issues, but the drawback is that the researcher must be very confident in the quality of the obtained solution.
2. If the solution does not fulfill all quality criteria, then the stratified bootstrap should be used. The advantage is that small components will not be excluded from the resampling procedure, but the drawback is that label-switching could occur and will have to be solved after each replication of the bootstrap procedure.
3. If the researcher is not sure about the suitability of the solution, then the non-parametric bootstrap should be applied, with the advantage of not fixing the structure of the clustering model, but with serious label-switching and multimodality issues possible.

It must be stressed that, when using either the separate or the stratified bootstraps, no inference can be made regarding the mixing proportion of the different components, and hence the importance for one to be very confident about the chosen model. Moreover, it must be remembered that, whatever the type of bootstrap, results are always conditional on the selected model, so the first critical point is the identification of the correct structure for the clustering model.

The importance of the chosen solution for the bootstrap procedures introduced in this study requires a discussion of the clustering choice and validation. One also needs to evaluate the impact of misidentification of the chosen model on both the separate and stratified bootstrap, as well

 ISSN: 2102-6238

as the importance of this impact compared to that of the label-switching problem and of the multimodality that occur when using the ordinary non-parametric bootstrap. These considerations could lead to other, alternative bootstrap procedures relaxing the fixed partition hypothesis used by the separate and stratified bootstrap procedures. For instance, the a-posteriori probabilities of component membership could be used to select the composition of the bootstrap samples. Further research on these questions is still required to evaluate the advantages and disadvantages of the different bootstrap procedures.

6. Acknowledgements

This publication benefited from the support of the Swiss National Centre of Competence in Research LIVES - Overcoming vulnerability: Life course perspectives, which is financed by the Swiss National Science Foundation (grant number: 51NF40-160590). The authors are grateful to the Swiss National Science Foundation for its financial support.

 ISSN: 2102-6238

TABLE 2. *Experiment 2: 95% CIs obtained for the HMTD model parameters using the three types of bootstrap.*

Situation 1: 200 sequences of length 25				
First component	σ_1^2	φ_{01}	φ_{11}	φ_{21}
True values	**4**	**5**	**0.7**	**0**
Non-parametric bootstrap	(1.861; 11.507)	(0.602; 5.538)	(-0.207; 1.034)	(-0.245; 0.769)
Separate bootstrap	(3.734; 4.626)	(2.387; 5.906)	(0.415; 0.995)	(-0.246; 0.342)
Stratified bootstrap	(1.879; 11.568)	(0.602; 5.570)	(-0.196; 1.009)	(-0.272; 0.775)
Second component	σ_2^2	φ_{02}	φ_{12}	φ_{22}
True values	**9**	**2**	**-0.2**	**0.6**
Non-parametric bootstrap	(0.852; 13.422)	(-0.480; 2.474)	(-0.268; 0.621)	(0.224; 0.841)
Separate bootstrap	(7.992; 9.992)	(1.612; 2.497)	(-0.273; -0.135)	(0.529; 0.656)
Stratified bootstrap	(1.056; 13.277)	(-0.442; 2.432)	(-0.266; 0.551)	(0.264; 0.845)
Third component	σ_2^2	φ_{02}	φ_{12}	φ_{22}
True values	**1**	**-0.5**	**0**	**0.3**
Non-parametric bootstrap	(0.313; 2.986)	(-0.681; 0.070)	(-0.100; 0.469)	(0.163; 0.423)
Separate bootstrap	(0.887; 1.112)	(-0.632; -0.393)	(-0.073; 0.070)	(0.219; 0.379)
Stratified bootstrap	(0.334; 2.952)	(-0.673; 0.074)	(-0.091; 0.572)	(0.185; 0.444)
Situation 2: 200 sequences of length 100				
First component	σ_1^2	φ_{01}	φ_{11}	φ_{21}
True values	**4**	**5**	**0.7**	**0**
Non-parametric bootstrap	(1.676; 10.656)	(0.575; 5.311)	(0.163; 1.042)	(-0.277; 0.784)
Separate bootstrap	(3.898; 4.590)	(2.505; 5.652)	(0.422; 1.014)	(-0.270; 0.336)
Stratified bootstrap	(1.453; 10.994)	(0.594; 5.361)	(0.167; 1.044)	(-0.271; 0.780)
Second component	σ_2^2	φ_{02}	φ_{12}	φ_{22}
True values	**9**	**2**	**-0.2**	**0.6**
Non-parametric bootstrap	(0.757; 13.124)	(-0.504; 2.210)	(-0.227; 0.173)	(0.276; 0.826)
Separate bootstrap	(8.526; 9.492)	(1.805; 2.221)	(-0.232; -0.170)	(0.568; 0.628)
Stratified bootstrap	(0.631; 12.939)	(-0.504; 2.211)	(-0.231; 0.167)	(0.250; 0.816)
Third component	σ_2^2	φ_{02}	φ_{12}	φ_{22}
True values	**1**	**-0.5**	**0**	**0.3**
Non-parametric bootstrap	(0.486; 2.860)	(-0.580; -0.444)	(-0.049; 0.045)	(0.254; 0.345)
Separate bootstrap	(0.937; 1.053)	(-0.557; -0.447)	(-0.038; 0.038)	(0.264; 0.336)
Stratified bootstrap	(0.449; 2.780)	(-0.602; -0.440)	(-0.080; 0.040)	(0.254; 0.343)
Situation 3: 800 sequences of length 25				
First component	σ_1^2	φ_{01}	φ_{11}	φ_{21}
True values	**4**	**5**	**0.7**	**0**
Non-parametric bootstrap	(1.598; 11.522)	(0.568; 5.398)	(-0.205; 0.998)	(-0.264; 0.783)
Separate bootstrap	(3.891; 4.509)	(2.680; 5.775)	(0.423; 0.986)	(-0.242; 0.325)
Stratified bootstrap	(1.756; 12.007)	(0.590; 5.430)	(-0.208; 1.019)	(-0.255; 0.776)
Second component	σ_2^2	φ_{02}	φ_{12}	φ_{22}
True values	**9**	**2**	**-0.2**	**0.6**
Non-parametric bootstrap	(0.496; 13.137)	(-0.502; 2.233)	(-0.230; 0.623)	(0.227; 0.842)
Separate bootstrap	(8.468; 9.510)	(1.775; 2.243)	(-0.233; -0.170)	(0.567; 0.631)
Stratified bootstrap	(0.684; 13.108)	(-0.484; 2.208)	(-0.235; 0.651)	(0.220; 0.841)
Third component	σ_2^2	φ_{02}	φ_{12}	φ_{22}
True values	**1**	**-0.5**	**0**	**0.3**
Non-parametric bootstrap	(0.302; 2.896)	(-0.594; 0.060)	(-0.066; 0.5100)	(0.223; 0.492)
Separate bootstrap	(0.946; 1.058)	(-0.561; -0.444)	(-0.039; 0.0355)	(0.259; 0.338)
Stratified bootstrap	(0.299; 2.939)	(-0.621; 0.052)	(-0.069; 0.4676)	(0.233; 0.416)

 ISSN: 2102-6238

References

Berchtold, A. (2003) Mixture transition distribution (MTD) modelling of heteroscedastic time series. *Computational statistics and data analysis* 41(3): 399-411.

Berchtold, A., & Raftery, A. (2002) The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science* 17(3): 328-356.

Berchtold , A., Suris, J. C., Meyer, T., & Taushanov, Z. (2018). Development of Somatic Complaints Among Adolescents and Young Adults in Switzerland. *Swiss Journal of Sociology*, 44(2): 239-257.

Bolano, D., & Berchtold, A. (2016) General framework and model building in the class of Hidden Mixture Transition Distribution models. *Computational Statistics & Data Analysis* 93: 131-145.

Celeux G., & Govaert G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28: 781-793.

Celeux, G., Hurn, M., & Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451): 957-970.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *The annals of Statistics*, 7(1): 1-26.

Grün, B., & Leisch, F. (2004). Bootstrapping finite mixture models. *Proceedings of the COMPSTAT 2004 Symposium*.

Jasra A., Holmes C., & Stephens D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, 20(1): 50-67.

Marin, J. M., Mengersen, K., & Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, Volume 25: 459-507.

Meila, M. (2016) Criteria for Comparing Clusterings, In Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.) *Handbook of cluster analysis* (Chapter 27). CRC Press.

O'Hagan,A., Murphy, T. B., Scrucca, L., & Gormley, I. C. (2018). Investigation of Parameter Uncertainty in Clustering Using a Gaussian Mixture Model Via Jackknife, Bootstrap and Weighted Likelihood Bootstrap. Available online at *https://arxiv.org/abs/1510.00551*

Raftery, A. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society, series B*, 47(3): 528-539.

Rodriguez, C. E., & Walker, S. G. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1): 25-45.

Rosychuk, R. J., Sheng, X., & Stuber, J. L. (2006). Comparison of variance estimation approaches in a two-state Markov model for longitudinal data with misclassification. *Statistics in medicine*, 25(11): 1906-1921.

Rydén, T. (2008) EM versus Markov chain Monte Carlo for estimation of hidden Markov models: a computational perspective. *Bayesian Analysis*, 3(4): 659-688.

Scrucca L., Fop M., Murphy T. B. & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1): 205-233.

Sperrin, M., Jaki, T., & Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20(3): 357-366.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, series B*, 62(4): 795-809.

Taushanov, Z., & Berchtold, A. (2017a). A Direct Local Search Method and its Application to a Markovian Model. *Statistics, Optimization & Information Computing*, 5(1): 19-34.

Taushanov, Z., & Berchtold, A. (2017b) Markovian-based Clustering of Internet Addiction Trajectories. In G. Ritschard & M. Studer (eds), *Sequence Analysis and Related Approaches: Innovative Methods and Applications*. Berlin: Springer.

Visser, I., Raijmakers, M. E., & Molenaar, P. (2000). Confidence intervals for hidden Markov model parameters. *British journal of mathematical and statistical psychology*, 53(2): 317-327.

 ISSN: 2102-6238